Niveau(x) d'enseignement: 0 Numéro EDUQ : 8670250
Centre d'intérêt: 540

Masson, Jean-Pierre.

Analyse et sélection des items pour un test critère / Jean-Pierre Masson . -- Sainte-Foy : IHRS-Education, [1973?].
34 f. ; 28 cm.

Bibliogr.: f. [32]-34. Code: R-86.

Vedette(s)-secondaire(s): IHRS-Education

Descripteur(s): Docimologie * Elaboration de tests * Revue de la documentation * Méthode d'analyse

Identificateur(s) de sujet: Test critère * Analyse d'item

L'analyse et la sélection des items constituent la deuxième étape du développement d'un test-critère, la première étant la construction même des items. Dans la première étape, la règle essentielle est de construire des items qui soient un reflet adéquat de l'objectif visé. Dans la deuxième étape, la règle essentielle est d'identifier un certain nombre d'items qui assurent une mesure fiable et permettent des prises de décision qui soient le plus juste possible. Le but de la recherche est de relever dans la documentation scientifique les différents indices proposés par les chercheurs pour l'analyse et la sélection des items du test-critère et de présenter certaines méthodes de construction d'items ainsi que certains modèles plus théoriques pouvant influencer grandement l'analyse et la sélection d'items.

Type de document: r

EDUQ 8670250

Analyse et sélection des items pour un test critère

Jean-Pierre Masson

INRS-Education

R-86

Université du Québec
INRS-Education
C.P. 7500, Sainte-Foy
Québec, G1V 4C7

INTRODUCTION

L'analyse et la sélection des items peuvent être considérées comme la deuxième étape du développement d'un test-critère, la première étape étant la construction elle-même des items. Dans la première étape, la règle essentielle est de construire des items qui soient un reflet adéquat de l'objectif visé. Dans cette deuxième étape, la règle essentielle est d'identifier un certain nombre d'items qui nous assurent une mesure fiable et nous permettent des prises de décision qui soient le plus juste possible.

Les méthodes d'analyse et de sélection d'items pour tests normatifs sont très bien développées mais, selon certains auteurs, sont difficilement applicables sinon inutilisables dans le cadre des tests-critères. Et ceci s'explique par le fait que ces deux types de test doivent conduire à des prises de décision de nature différente. Les docimologues doivent donc soit créer de nouvelles méthodes, soit reformuler des méthodes déjà existantes, soit réinterpréter ces méthodes dans un nouveau contexte.

Le but du présent travail est de relever dans la littérature scientifique les différents indices proposés par les chercheurs pour l'analyse et la sélection des items de test-critère. A cela cependant, nous ajouterons une présentation de certaines méthodes de construction d'items de même que certains modèles plus théoriques pouvant influencer grandement l'analyse et la sélection d'items. Nous craindrions qu'effectuer une simple présentation d'indices donne une idée inappropriée des recherches qui se font dans ce domaine.

SELECTION ET METHODES DE CONSTRUCTION

Plusieurs méthodes de construction d'items pour tests-critères impliquent dans leur fondement même la sélection d'items selon des règles précises. Klein et Kosecoff (1973) classent les différentes méthodes existantes en trois catégories. Dans une première catégorie, c'est un groupe d'experts, se fiant à leurs connaissances et à leur expérience, qui décident des items à utiliser; il n'y a donc pas ici de règles précises de sélection à ce niveau de construction. La deuxième catégorie implique comme approche le développement pour chacun des objectifs d'une matrice de contenus et comportements. Le choix des items s'effectue alors selon un échantillonnage systématisé à l'intérieur de la matrice. La troisième catégorie constitue l'approche la plus sophistiquée;on construit des formules de base qui vont produire les items nécessaires par des techniques précises.

Cette dernière approche doit certainement être décrite ici parce qu'elle oblige à considérer l'analyse d'items sous un angle très particulier. Elle peut même mettre son existence en cause. En effet, cette approche définit complètement tous les contenus pertinents, les comportements ou tâches désirés de même que les stimuli et les caractéristiques de la réponse. Aussi Ivens, dans une revue de la littérature portant sur l'analyse d'items la traite indépendamment.

> In constructing criterion-referenced measures from the item-form approach, item difficulty as such is an irrelevant concept (Ivens, 1970, page 7).

> In the item-form approach to criterion-referenced testing, item homogeneity within a given objective is assured in that all items must be parallel by the very nature of the item form. Similar considerations apply to the content validity of the items (Ivens, 1970, page 9).

Cette approche est représentée principalement par deux auteurs: Bormuth et Hively. Nous allons décrire la pensée de chacun.

Hively, Patterson et Page

Hively, Patterson et Page présentent l'une des méthodes les plus sophistiquées pour la construction et la sélection d'items de tests-critères. Leur méthode s'appuie sur le behaviorisme qui démontre qu'il est possible d'analyser une matière d'études en classes de comportement bien définies, et sur la théorie de la généralisabilité de Cronbach. Cronbach postule qu'un test de performance est un échantillon d'un univers d'items large mais bien défini; des tests parallèles sont obtenus par un échantillonnage répété selon un plan spécifié. Osburn, en 1969, avait décrit la valeur théorique des modèles de mesure auxquels cette théorie pouvait donner naissance et a nommé ce domaine "universe-defined achievement testing". La méthode de Hively, Patterson et Page en est la première application pratique.

Leur travail a porté sur les mathématiques. Ils ont divisé la matière en parties significatives en se référant aux catégories développées par les enseignants dans la construction de tests diagnostics. Ils définissent de façon précise l'ensemble de tous les items de test qui peuvent être pris pour représenter une catégorie en élaborant une formule générale et une liste de règles génératives.

Les règles pour générer un tel ensemble d'items ont appelées "formules d'items". Une collection de formules d'items constitue un "univers" duquel on peut tirer tous les tests. On parle de l'univers de l'addition, l'univers de la soustraction, etc.

Une famille de tests parallèles et randomisés est définie par un plan d'échantillonnage portant sur un univers de formules d'items. On utilise les termes "famille de l'addition", "famille de la soustraction" pour référer à l'ensemble de tous les tests possibles pouvant être constitués par un tel échantillonnage.

> Currently, the most useful scheme for writing items which possess congruency with a criterion has been provided by Hively and his associates (Hively, W., 1970) in their work on *domain-referenced achievement testing*. In this system an *item form* constitutes a complete set of rules for generating a domain of test items which can be used to accurately measure a particular objective (Popham, 1971, page 81).

Bormuth

Dans son volume "Theory of Achievement Test Items", Bormuth présente une approche opérationnelle à la construction d'items de test de performance. Le postulat fondamental de cette approche pose qu'un item de test de performance ne peut pas produire de résultats significatifs à moins qu'il ne soit relié à et dérivé de l'enseignement par un ensemble d'opérations. Les opérations elles-mêmes décrivent un ensemble de manipulations directement observables par lesquelles un item de test est dérivé de l'enseignement.

Pour Bormuth, l'item est le produit de définitions opérationnelles. Une définition opérationnelle est une description de la

façon dont une classe d'items est dérivée de l'enseignement. Elle a deux composantes majeures: un ensemble d'opérations par lesquelles une structure syntaxique est assignée à l'enseignement et un ensemble d'opérations effectuées sur cette syntaxe qui transforment les segments pertinents de l'enseignement en items de test. En d'autres mots, l'enseignement lui-même est transformé en questions et réponses par des manipulations commandées par les opérations constituant la définition.

Un exemple de cette procédure: les questions de type "suppression du sujet". Cette procédure consiste en deux séries d'opérations. La première spécifie la procédure pour identifier les sujets de toutes les phrases contenues dans l'enseignement. La deuxième série d'opérations demande au rédacteur de supprimer tous les sujets et de remplacer chacun par le pronom relatif approprié: qui, que, quoi, où. La réponse correcte est alors soit l'élément supprimé ou un dérivé de cet élément, dérivé obtenu par d'autres opérations. Par exemple, si nous avons la phrase "Le fermier laboure son champ", nous obtenons la question "Qui laboure son champ?" et aucune autre question pour ce segment de l'enseignement ne peut être formulée sans que d'autres définitions ne soient construites.

Selon Dornuth, cette approche pour la rédaction d'items est la seule qui permette une méthode de construction véritablement objective puisqu'elle permet d'expliquer entièrement le processus de rédaction d'item et ne laisse aucune place à la subjectivité du rédacteur. Deux rédacteurs de test utilisant les mêmes définitions et le même enseignement, mais travaillant indépendamment devraient produire des items similaires jusque dans les moindres détails.

ANALYSE D'ITEMS

Klein et Kosecoff (1973) ont identifié deux concepts de base sous-jacents aux diverses techniques d'analyse d'items proposées dans le cadre des tests-critères. Au moins un de ces concepts serait présent dans chacune des méthodes d'analyse.

Le premier concept se présente comme la sensibilité de l'item à l'enseignement. Si le rendement à l'item est relié au degré d'enseignement effectué, l'item est considéré "bon". Le deuxième concept se présente comme la discrimination de l'item entre ceux qui réussissent bien versus ceux qui réussissent mal à l'ensemble du test ou à un critère extérieur, tel le rendement à la séquence d'enseignement suivante.

Ces auteurs regroupent les diverses méthodes d'analyse en quatre types qu'ils présentent ainsi: Le premier type réfère aux méthodes exigeant la passation du même test à deux groupes reconnus pour posséder un degré de compétence différent relativement à l'objectif mesuré. Une façon de faire est d'administrer le test à ceux qui ont reçu un enseignement sur l'objectif et à ceux qui n'en ont pas reçu. On identifie alors les items qui discriminent bien entre les deux groupes.

Le deuxième type réfère aux méthodes requérant la passation du même test au même groupe à deux reprises, une avant l'enseignement et une autre après l'enseignement. On identifie alors les items qui discriminent bien entre les deux sessions de test. Parmi les auteurs qui utilisent cette approche, nous retrouvons Popham (1970), Ozenne (1971), Kosecoff et Klein (1973), et Roudabush (1973).

Dans un troisième type, on soumet le test à un groupe d'étudiants après une certaine période d'enseignement. Tous les étudiants ont alors eu les mêmes possibilités pour atteindre l'objectif. Si le temps alloué est moins que celui dont tous les étudiants avaient besoin pour atteindre l'objectif et si les étudiants diffèrent dans leurs habiletés, alors on utilise les méthodes d'analyse classiques pour identifier les items à rejeter ou à conserver. Comme critère on peut utiliser le score total au test ou le rendement à une habileté plus avancée.

Dans un quatrième type, chaque étudiant passe périodiquement le test au complet jusqu'à ce qu'il atteigne la maîtrise de l'objectif. On enregistre le nombre de fois que l'étudiant réussit et échoue chacun des items. L'analyse a pour but de déterminer si l'item se comporte selon un schéma désiré. Par exemple, un schéma souhaitable serait Echec - échec - réussite - réussite, alors qu'un schéma indésirable serait Réussite - échec - réussite - échec. Plusieurs conditions doivent être respectées pour que cette méthode puisse être appliquée.

Des indices

Ivens, Cox, Vargas

Selon Ivens, un indice aussi simple que la différence entre le niveau de difficulté de l'item au post-test et le niveau au pré-test est déjà valable pour le choix d'un item. Idéalement, un item de test-critère aurait une cote de difficulté "zéro" (ou du niveau de la chance) au pré-test et une cote "un" au post-test. Cela démontrerait la nécessité et l'efficacité de l'enseignement. Cox et

Vargas (1966) ont comparé cet indice avec l'indice de difficulté utilisé traditionnellement dans les tests normatifs et n'ont trouvé qu'une corrélation de .37 dans un cas et de .40 dans un autre. Une étude de Popham confirme ces résultats. Ils concluent ainsi que l'indice de la différence pré-test - post-test permet de choisir des items bien différents de ceux choisis par les indices classiques, ceux-ci n'étant basés que sur le potentiel discriminatoire des items lors d'un post-test seulement.

Ivens suggère deux autres indices pour juger de la valeur d'un item. Ces deux indices exigent trois administrations différentes du même test aux mêmes sujets: pré-test, post-test et re-test. Si le test se comporte comme souhaité, les scores au pré-test se maintiendraient autour d'un niveau de probabilité (chance); ils se maintiendraient près du niveau de maîtrise au post-test et au re-test. Autrement dit, pour chaque item et pour chaque sujet, nous obtiendrions un changement maximal dans la performance entre pré-test et post-test, et un changement minimal entre post-test et re-test.

Le premier indice, déjà suggéré par King (1970), consiste à calculer la valeur de l'expression suivante:

(1 - concordance pré-post-test) (concordance post-re-test)

La concordance se définit comme la proportion des sujets dont le score à l'item (succès ou échec) est identique pour les deux administrations. Cet indice atteindra une valeur maximale de un (1) si chacun des sujets manque l'item au pré-test et passe l'item au post-test et au re-test. Inversement, la valeur maximale de zéro(0) sera atteinte si chacun des sujets passe (ou manque) l'item au pré-test et au post-test, ou si chacun des sujets obtient des résultats différents à l'item au post-test et au re-test.

Le deuxième indice consiste à calculer pour chaque item la valeur de l'expression:

$$(p_{post} - p_{pré})\ (1 - /p_{re\text{-}test} - p_{post}/)$$

p représente la proportion de sujets passant un item donné à une administration donnée. Cet indice, bien que similaire au premier, repose sur des données de groupe alors que le premier repose sur des données individuelles. Ce deuxième indice peut aller d'une valeur maximale de un (1) à une valeur minimale de moins un (-1). Des valeurs inférieures à zéro ne peuvent être atteintes que si la proportion des sujets passant l'item au pré-test est plus grande que la proportion le passant au post-test.

L'homogénéité des items construits pour un objectif donné peut aussi servir lors de la sélection des items. Cette homogénéité peut être évaluée au moyen des cotes de difficulté du pré-test et du post-test. L'ensemble des items portant sur un objectif donné est utilisé comme le contrôle ou l'étalon vis-à-vis lequel chaque item est évalué. Ivens suggère d'analyser soigneusement tout item d'un groupe qui présente un indice de difficulté vraiment différent des autres.

Hambleton, Gorth, Wardrop

La suggestion d'Ivens est aussi présentée par Hambleton et Gorth (1971).

> Therefore all of the items associated with the same objective should be answered correctly by about the same proportion of examinees in a group, i.e., they should have approximately the same value for the item difficulty index. If an item

> has a value of the index quite different from all of the other items, it probably is measuring a performance which is identifiably different from the objective (Hambleton et Gorth, 1971, page 7).

Par contre, différemment de Ivens, Hambleton et Gorth calculent l'indice de difficulté selon les méthodes traditionnelles pour les tests normatifs. Si l'indice des items associés à un objectif diffère, deux alternatives sont possibles: ou bien les items qui ressemblent le moins à l'objectif devraient être modifiés, ou bien l'objectif devrait être redéfini plus spécifiquement afin de référer à des items similaires portant des indices similaires.

Hambleton et Gorth (1971) suggèrent aussi d'utiliser les indices de discrimination développés pour les tests normatifs. Des indices de discrimination négatifs indiquent que des items inclus dans le test-critère doivent être modifiés ou quelquefois, que l'enseignement ou le matériel pédagogique utilisé a été inefficace. Rappelons qu'une valeur négative signifie que les étudiants qui ont généralement le mieux réussi au test total ont répondu incorrectement à l'item plus fréquemment que les étudiants qui ont faiblement réussi.

Cependant, de l'avis même de Hambleton et Gorth, la suggestion la plus intéressante aurait été formulée par Wardrop (1970). Celui-ci propose comme alternative aux procédures traditionnelles d'analyser une courbe caractéristique de l'item. Cette courbe a été étudiée dans le cadre d'un modèle des traits latents, modèle qui spécifie une fonction reliant la probabilité de succès à un item à l'habileté ou trait latent sous-jacent mesuré par le test.

L'habileté ou trait latent pour chacun des individus peut être conceptualisé comme la position de l'individu sur une échelle d'habileté allant d'une performance nulle à une performance parfaite.

Et le problème de la mesure est de situer l'individu au bon endroit sur ce continuum. Parmi les diverses fonctions proposées, la courbe à deux paramètres de Brinbaum (1968) est à retenir.

$$P_g(x) = \left(1 + e^{-Da_g (x - b_g)}\right)^{-1}$$

Cette courbe serait caractéristique d'un item. $P_g(x)$ est la probabilité qu'un sujet avec une habileté $\underline{x}$ réponde correctement à l'item $\underline{g}$. Le paramètre $\underline{b_g}$ réfère à l'indice de difficulté de l'item, alors que $\underline{a_g}$ réfère à l'indice de discrimination. D est une constante.

Quel est l'intérêt de cette approche ? D'abord, en théorie du moins, les paramètres "difficulté" et "discrimination" ne varient pas d'un groupe à l'autre ce qui est généralement faux pour les paramètres d'item traditionnels. L'indice de difficulté classique, défini comme la proportion de sujets dans un groupe qui répond correctement à l'item, varie en fonction de l'habileté du groupe. L'invariance du paramètre "difficulté" permettrait la construction de tests à caractéristiques spécifiques sans connaissance préalable de l'habileté des sujets. Cependant, cette approche exige le développement d'une procédure adéquate d'estimation des paramètres et des habiletés de même que la vérification de l'invariance des paramètres. Ces problèmes ne sont pas encore résolus.

Tse-Chi-Hsu

Selon Tse-Chi-Hsu (1971), il n'existe pas encore, à cette date, d'indices statistiques valables qui permettent de sélectionner les items de test-critère. La citation qui suit résume sa critique et présente l'essentiel de sa pensée.

> If the test is going to be used to provide "explicit information as to what the individual can or cannot do" (Glaser, 1963), then a good criterion-referenced test item does not only discriminate pre- and post-learning. It is also the function of the item to allow the individual to answer correctly if he masters the criterion behavior represented by the item and answer incorrectly if he actually does not master it, regardless of whether the test is administered before or after formal instruction (Tse-Chi-Hsu, 1971, page 3).

Pour Tse-Chi-Hsu, il faut d'abord diviser une population en deux groupes, celui qui maîtrise le comportement-critère et celui qui ne le maîtrise pas. Cette classification est faite sur la base du résultat au test. Alors, pour chacun des items, il trouve la proportion de sujets dans le groupe-maîtrise qui ont répondu correctement à l'item et la proportion de sujets dans le groupe-non-maîtrise qui ont aussi répondu correctement. Selon lui, la différence de ces deux proportions (D_p) constitue un indice de discrimination valable pour les tests-critères. Pour lui, une autre façon de calculer ce type de discrimination serait d'utiliser le coefficient phi ϕ. Un tableau à double entrée devrait être utilisé: réponse correcte (1), et réponse incorrecte (0); maîtrise (1) et non-maîtrise (0). Ce coefficient serait, à son avis un meilleur indice que la corrélation tétrachorique puisque celle-ci doit assumer une distribution bivariée normale.

L'auteur a aussi étudié la corrélation pouvant exister entre les indices D_p et ϕ de même qu'avec la corrélation point-bisériale (r_pbi). Les résultats indiqueraient que dans la plupart des cas, ces indices sont fortement corrélés.

Dahl

Toute l'approche de Dahl est centrée sur la notion de congruence, entre objectif et items de test. Pour lui, les techniques d'analyse devraient regrouper de quelque façon les items qui sont congruents à un objectif de telle manière qu'ils se distinguent de tous les autres groupements d'items. Malheureusement pour lui, les techniques habituelles d'analyse de groupements et d'analyse de facteurs ne sont pas applicables aux types de données recueillis par les tests-critères. Dahl propose alors une procédure connue sous le nom d'analyse de monotonicité. Cette technique a été développée par Bentler (1968) et pourrait constituer une méthode valable pour mesurer la congruence. Cependant, on ne possède pas encore de données permettant de vérifier la validité de cette assertion.

Popham

Popham (1971), après avoir utilisé diverses procédures d'analyse d'items, arrive à la solution suivante. Selon lui, ce qu'il s'agit de faire, c'est d'identifier les items qui se comportent différemment des autres, items mesurant le même objectif, ou, dit en d'autres mots, d'un item-prototype. Une façon de faire serait de rassembler d'abord les données recueillies lors d'un pré-test et d'un post-test de la façon suivante.

	01	00	11	10
Item 1				
Item 2				
Item 3				
Etc.				

0: échec
1: succès

Il s'agirait alors de calculer la valeur médiane de chacune des colonnes. Pour chacun des items, les fréquences obtenues dans les cellules (01, 00, 11, 10) seraient ensuite comparées aux fréquences hypothétiques basées sur les valeurs médianes par un chi carré. La question qu'on devrait alors se poser est la suivante: comment préciser la limite minimum à atteindre pour les valeurs du chi carré ?

Roudabush

Roudabush considère que les statistiques classiques pour analyser des items demeurent utiles pour étudier un test. Cependant, il considère plus important de déterminer si les items sont sensibles à l'enseignement. Afin d'obtenir un indice numérique réflétant la sensibilité de chacun des items, il administre le test avant et après l'enseignement, et organise les données recueillies selon la table de fréquences suivante:

Pré-test \ Post-test	0	1	
0	f_1	f_2	$f_1 + f_2$
1	f_3	f_4	$f_3 + f_4$
	$f_1 + f_3$	$f_2 + f_4$	N

f_1 : la fréquence des sujets ayant échoué l'item et au pré-test et au post-test.

f_2 : la fréquence des sujets ayant échoué l'item au pré-test, mais l'ayant réussi au post-test.

f_3 : la fréquence des sujets ayant réussi l'item au pré-test, mais l'ayant échoué au post-test.

f_4 : la fréquence des sujets ayant réussi l'item et au pré-test et au post-test

N : $f_1 + f_2 + f_3 + f_4$ = le nombre total de sujets ayant reçu l'administration de l'item et au pré-test et au post-test.

L'auteur appuie son rationnel sur quatre postulats.

1. Il existe une probabilité (p) fixe et différente de zéro qu'un sujet qui ne connaît pas la réponse devine la réponse correcte. La valeur de p est déterminée seulement par l'item et ne varie pas d'un sujet à l'autre, ni d'une occasion à l'autre pour un étudiant.

2. Les réponses d'un sujet sont indépendantes au pré et au post-test quand il ne connaît pas la réponse correcte et échoue à l'apprendre: il n'y a pas de connaissance partielle.

3. Le seul résultat possible de l'apprentissage entre le pré- et le post-test est que le sujet apprenne la réponse correcte à l'item.

4. Il n'existe pas d'oubli de sorte qu'une fréquence f_3 différente de zéro n'est due qu'au hasard (guessing). La valeur "vraie" de f_3 est zéro.

Le rationnel suivi par Roudabush pour obtenir un indice numérique de sensibilité se présente comme suit: La fréquence (f_1) des sujets qui ont échoué l'item et au pré-test et au post-test ne comprend que

ceux qui de fait ne connaissent pas la réponse après apprentissage. Ainsi, f_1 est égale à la probabilité pour le nombre de sujets dans l'échantillon d'opter deux fois pour une mauvaise réponse. Donc,

$$f_1 = (1 - p)^2 Fc_1 \qquad (1)$$

Fc_1 représente le nombre "vrai" de sujets qui ne savent pas. De la même façon, la fréquence (f_2) de ceux qui ont échoué l'item au pré-test et l'ont réussi au post-test, ne représente que le nombre de sujets qui ont appris la réponse correcte et ont opté pour une mauvaise réponse au pré-test, plus le nombre de sujets qui n'ont pas appris mais qui ont deviné correctement au post-test et ont opté pour une mauvaise réponse au pré-test. Donc,

$$f_2 = (1 - p) Fc_2 + p (1 - p) Fc_1 \qquad (2)$$

Fc_2 représente le nombre "vrai" de sujets dans l'échantillon qui ne savaient pas au pré-test mais ont appris pour le post-test.

La fréquence (f_3) de ceux qui ont réussi l'item au pré-test mais l'ont échoué au post-test n'est encore composée que de ceux qui ne connaissent ni n'ont appris la réponse correcte mais qui ont deviné correctement au pré-test. Donc,

$$f_3 = p (1 - p) Fc_1 \qquad (3)$$

Finalement, la fréquence (f_4) de ceux qui ont réussi l'item et au pré-test et au post-test représente (1) tous ceux qui de fait reconnaissaient la réponse correcte et au pré-test et au post-test, (2) tous ceux qui ont appris la réponse correcte et ont deviné la bonne réponse au pré-test, et (3) tous ceux qui ne connaissaient pas la bonne réponse et ne l'ont pas appris mais qui ont deviné correctement au pré-test et au post-test. Donc,

$$f_4 = Fc_4 - pFc_2 - p^2Fc_1 \qquad (4)$$

Fc_4 représente le nombre "vrai" de ceux qui dans l'échantillon connaissaient la réponse correcte et au pré-test et au post-test.

En partant des équations (1) et (3), nous obtenons,

$$p = \frac{f_3}{f_1 - f_3} \qquad (5)$$

et les équations (1) à (4) forment un système cohérent de sorte que l'on peut trouver les solutions pour Fc_i

$$Fc_1 = \frac{(f_3 - f_1)^2}{f_1}$$

$$Fc_2 = \frac{(f_2 - f_3)\ (f_2 - f_1)}{f_1} \qquad (6)$$

$$Fc_3 = 0$$

$$Fc_4 = f_4 - \frac{f_3\ f_2}{f_1}$$

Le rapport

$$I.S. = \frac{Fc_2}{Fc_1 - Fc_2} \qquad (7)$$

peut servir comme indice du degré auquel les sujets choisissent la bonne réponse à l'item en fonction de l'apprentissage effectué entre le pré- et le post-test. C'est l'indice de sensibilité. Cet indice n'est simplement que la proportion de sujets ayant échoué l'item au pré-test et l'ayant réussi au post-test après qu'une correction pour le hasard aie été effectuée.

QUELQUES MODELES D'ANALYSE DE TESTS-CRITERES

Dans ce chapitre, nous vous présentons quatre modèles différents d'analyse de tests-critères. Ces modèles dépassent le niveau d'une simple analyse d'items mais la renferment, implicitement ou explicitement en nous offrant un cadre d'analyse plus global. Pour le moins, ils permettent de mieux se situer face à l'analyse d'items.

Le modèle de Besell

Plusieurs auteurs, parmi lesquels Livingston, Block, Airasian, Madaus, etc..., ont déclaré que l'interprétation de la performance d'un étudiant ne devrait pas être dépendante de la performance de ses pairs. Besell ne partage pas cet avis et croit au contraire que la performance d'un groupe normatif fournit des informations fort utiles à la construction de tests-critères. Son modèle d'analyse, appelé "Mastery Learning Test Model" se sert de ces informations comme de données essentielles. Il postule que chaque étudiant d'un groupe d'apprentissage peut être considéré comme appartenant à l'une de ces deux catégories: celle qui a maîtrisé l'objectif et celle qui n'a pas maîtrisé l'objectif. Se référant au modèle d'Emrick (1971), il postule que l'erreur de mesure peut être expliqué par les deux paramètres suivants:

A) la probabilité qu'un non-maître donne une réponse correcte à un item,

B) la probabilité qu'un maître donne une réponse incorrecte à un item.

Ces paramètres sont évalués à partir de données recueillies chez un échantillon représentatif d'étudiants. Selon ce modèle, les seuls scores vrais peuvent être zéro (0) et N (pour un test comprenant N items); tous les scores intermédiaires sont considérés comme dus à une erreur de mesure. La probabilité que tel étudiant ait atteint l'objectif et la proportion d'étudiants dans le groupe d'apprentissage qui ont atteint l'objectif constituent les deux autres statistiques importantes.

Puisque le modèle de Beseil s'appuie sur le modèle statistique de Bayes, on doit d'abord effectuer une première estimation de la dernière probabilité mentionnée. On peut utiliser directement la proportion d'étudiants considérés comme ayant atteint l'objectif: on assigne alors la même probabilité à chaque étudiant; lorsque la moyenne du groupe augmente, cette première estimation de la probabilité augmente aussi pour chacun des membres du groupe. Il est cependant possible d'individualiser cette probabilité si on possède des données pour chacun des individus sur d'autres tests.

On effectue alors une analyse séquentielle des réponses aux items, selon une procédure définie. Cette procédure est répétée pour chacun des items, la probabilité de maîtrise résultant du premier item étant utilisée comme la probabilité de base du second item et ainsi de suite.

Ce modèle, au moyen d'un ensemble d'équations et de formules vous conduit à une estimation de la probabilité que tel étudiant ait atteint tel objectif représenté dans le test par un ensemble d'items. Il s'agit d'un indice de confiance.

L'approche d'Hambleton et Novick

Hambleton et Novick envisagent le problème des tests-critères

comme celui de séparer les étudiants en deux groupes distincts: un groupe constitué de ceux dont le résultat au test est assez élevé pour inférer qu'ils ont maîtrisé le matériel au niveau de rendement désiré et ceux qui n'ont pas atteint le minimum de rendement désiré.

Le premier problème est de déterminer si le véritable niveau de maîtrise de l'étudiant (Πi) est plus grand que le standard établi (Πo). Πi représente la proportion d'items d'un contenu particulier auquel l'étudiant peut répondre. Mais puisqu'on ne peut administrer tous les items du domaine investigué, nous utilisons un échantillon de ces items pour obtenir une estimation de Πi, soit $\hat{\Pi i}$. La valeur Πo, au présent stade de développement des tests-critères, est un score arbitraire servant de seuil pour diviser les individus dans les deux groupes mentionnés plus haut.

Le problème de l'examinateur est de placer l'étudiant dans la bonne catégorie. Deux types d'erreur peuvent se produire: une erreur fausse-positive a lieu lorsque l'examinateur situe l'examiné au-dessus du seuil alors qu'il ne l'est pas; une erreur fausse-négative a lieu quand l'examinateur situe l'examiné au-dessous du seuil alors que l'inverse est vrai.

Le point central de l'approche de Hambleton et Novick réside dans la notion de "perte". Par exemple, la perte ou le dommage causé par une erreur fausse-positive (type a) dépend de la structure des objectifs pédagogiques; lorsque les objectifs sont de structure hiérarchique, les conséquences sont sérieuses. La perte causée par une erreur fausse-négative (type b) dépend du temps pendant lequel l'étudiant sera assigné à une tâche de récupération. On peut aussi considérer d'autres facteurs tels que le coût du matériel, le temps du professeur, etc. Le professeur n'a pas à spécifier la valeur absolue (grandeur de la perte) de a et de b, mais il doit préciser le rapport a/b. Ce rapport servira à déterminer un seuil de perte et

des formules statistiques sont développées pour assurer des décisions qui engendront le moins de perte possible.

Kriewall

Selon Kriewall, le concept de difficulté d'item, pris dans un sens classique, n'est pas approprié aux tests-critères. Les procédures classiques de sélection d'items ne peuvent servir que de guide et non d'algorithme pour la construction de test. L'approche de Kriewall part de la définition suivante:

> A learning objective (1°) is a rule for generating a class of performance tasks, or alternatively a list of all performance tasks which comprise the objective (Kriewall, 19 p.)

Pour lui, la valeur d'un test-critère repose sur l'application d'un modèle d'échantillonnage d'item très strict; par très strict, il signifie que l'on doit d'abord définir la population d'items, et ensuite choisir au hasard un échantillon de N items pour le test. Et il postule que la compétence d'un individu, à un moment précis, est égale pour toutes les tâches d'une classe donnée. En d'autres mots, tous les problèmes contenus dans une classe sont d'égale difficulté pour l'individu. Il postule de plus que le rendement d'un individu à un item ou tâche donnée est indépendant de son rendement à tout autre item du même test-critère. Kriewall considère donc la performance à un test-critère comme une séquence d'essais indépendants, chaque essai ayant la même probabilité de succès.

Selon lui, le besoin de l'enseignant est de connaître, à un moment donné, la probabilité de succès que possède un individu face à une catégorie spécifique de tâches. Il assume qu'un individu possède une compétence unique face à un objectif précis. Comme mesure

de cette compétence, il prend le score relatif vrai de l'individu sur une population d'items. La compétence devient un paramètre qui peut être interprété comme la probabilité qu'un individu donne une réponse acceptable à n'importe quel item tiré au hasard de la population spécifiée.

Le but d'un test-critère devient alors l'estimation de la difficulté d'ensemble d'une classe d'items pour un individu. De la même façon que la valeur Π (score relatif attendu à un item pour une population d'examinés) sert de mesure de la difficulté d'un item pour une population d'élèves, ζ_q est considéré comme la difficulté moyenne d'une population d'items pour un individu donné.

L'une des implications de ce modèle porte sur la modalité des items. Afin de réduire l'erreur due à des biais, les items ne doivent pas être du type choix multiple puisque dans ce cas, le score attendu quand $\zeta = 0$ n'est pas $= \zeta$ mais est déterminé par le nombre de distracteurs utilisés. Les items à réponse ouverte sont hautement désirables bien qu'ils soient encore aujourd'hui la source de problèmes d'ordre économique.

Livingston

Livingston s'est efforcé d'adapter la théorie des tests classiques aux tests-critères. L'essentiel de son approche consiste à remplacer le concept de moyenne par le concept de score-critère et ensuite d'adapter les formules classiques à cette substitution.

Selon Livingston, la différence de base entre une mesure normative et une mesure critériée réside dans le fait que, pour des mesures normatives, nous voulons savoir à quelle distance un étudiant se situe par rapport à la moyenne du groupe, alors que pour des mesures critériées, nous voulons savoir à quelle distance il se situe

d'un standard fixé. Il conclue ainsi que chacun des concepts basés sur les déviations à la moyenne pourrait être remplacé par un concept correspondant basé sur les déviations ou score critère.

Cette approche a été sévèrement critiquée par plusieurs auteurs dont Harris.

SYNTHESE

Nous constatons, à la lumière des pages précédentes, que l'analyse d'items ne peut être conçue indépendamment des autres niveaux d'analyse ou de construction. En effet, à l'étape même de construction des items, des décisions doivent être prises qui ont des conséquences directes sur l'analyse d'items.

> There is some evidence that tests constructed from carefully defined domains of items possess reasonably good psychometric properties without prior statistical selection (e.g. Ebel, 1962, Hively, Patterson, and Page, 1968) (Nilko, 1971, page 8)

Par contre, il est assez difficile de respecter les exigences d'une telle méthode de construction et le plus souvent, les items sont construits ou bien selon des matrices contenus-comportements ou bien sont dérivés directement de l'objectif pédagogique. L'analyse d'items doit tenir compte de la méthode utilisée.

Au niveau même de l'analyse d'items, plusieurs méthodologies différentes peuvent être utilisées. Les données peuvent être recueillies lors d'un pré-test et d'un post-test à un même groupe de sujets ou bien par un post-test à deux groupes possédant des habiletés différentes, ou bien par la passation périodique, en cours d'apprentissage, du même test à un groupe d'étudiants ou finalement par la passation d'un test après une période critique d'apprentissage. Evidemment, selon la méthodologie utilisée, nous pourrons obtenir des indices de nature différente sur la qualité de nos items. Il existe ainsi des indices de difficulté, des indices de discrimination, des indices de sensibilité, des mesures de congruence, des

comparaisons à une courbe caractéristique, des comparaisons à des fréquences théoriques possédées par un prototype d'item.

Pour au moins la moitié de ces indices, il existe plusieurs formulations et autant de méthodes de calcul. Il est difficile d'établir celle qui est la plus valable.

Dans un tel puzzle, pour procéder de façon rationnelle et cohérente, il est certainement préférable de se référer à un modèle d'analyse docimologique plus complet. Or plusieurs modèles sont suggérés présentant des conceptions passablement différentes des tests-critères. Encore là un choix doit être fait. Et ce qui complique le travail du docimologue, c'est que ces modèles, pour la plupart, sont, pour l'instant incomplet. Il est donc très difficile présentement d'établir l'analyse et la sélection d'items pour tests-critères sur des bases solides, et de valeur assurée.

CRITIQUE

Que doit on retenir des diverses méthodes d'analyse d'items proposées jusqu'à maintenant ? La question est bien concrète car c'est celle que se pose nécessairement le docimologue qui a la responsabilité de construire un test-critère. Je ne crois pas qu'il existe une réponse unique à cette question. Cependant, certaines considérations peuvent recevoir l'assentiment de la plupart des gens travaillant dans ce domaine.

Parmi des considérations, la première et peut-être la plus importante a été soulevée par Nitko (1971).

> Elimination of items on purely statistic ground is poor practive generally in achievement test construction and becomes increasingly serious in criterion-referenced testing (Nitko, 1971, page 8).

En effet, lors de l'élaboration des items, le docimologue vise à représenter dans son test, et ce le plus adéquatement possible, le domaine de tâches qu'il veut investiguer. En se fiant uniquement à des données statistiques pour sélectionner les items qui constitueront finalement le test, il y a danger d'invalider ou de biaiser les résultats, ou tout simplement de ne plus recouvrir adéquatement le domaine de tâches en question. Il ne faut jamais oublier que la qualité essentielle d'un test critère est sa validité de contenu.

> Above all else, a criterion-referenced test must have content validity (Hambleton et Gorth, 1971, page 14).

Ceci signifie que les décisions à prendre concernant la rétention ou le rejet d'un item ne doit pas se faire uniquement sur des bases statistiques. Les statistiques fournissent certes des informations pertinentes à la prise de décision mais elles ne doivent qu'aider le docimologue dans sa réflexion et non engendrer des automatismes de réponse. Dans cette perspective, nous pouvons dire que plus le docimologue possèdera d'informations, c'est-à-dire que plus le nombre d'indices calculés sera grand, meilleure sera la vision de celui-ci sur ses items. Il est évident toutefois que des facteurs d'économie entreront ici en jeu. Il est évident aussi que la particularité des situations fera que le docimologue accordera plus ou moins de crédit à tel ou à tel autre indice statistique. L'expérience du docimologue aura ici sa place.

Dans sa réflexion, le docimologue aura aussi à tenir compte d'un autre facteur, soit le contexte pédagogique dans lequel se situeront les tests qu'il aura construits. En effet, il est important que l'appareil docimologique s'intègre bien au modèle pédagogique où il s'inscrit. Les deux ne peuvent être considérés indépendamment. Qu'on pense simplement aux exigences particulières que posent les divers modèles d'enseignement individualisé. Donc, dépendamment de l'utilisation qu'on fera des résultats produits par les tests, certaines méthodes d'analyse d'items devront être privilégiés par rapport à d'autres.

Finalement, la méthode de construction utilisée nous amènera à mettre plus ou moins d'emphase sur les analyses de type statistique. En ce sens, l'application des méthodes telles que proposées par Hively ou par Bormuth peuvent justifier la mise en second plan des méthodes d'analyse d'items.

Jusqu'ici, nous avons discuté de l'utilisation des méthodes d'analyse d'items proposés. Maintenant il serait bon d'émettre

quelques opinions sur ces méthodes. Ces opinions sont très personnelles et font certainement appel à beaucoup de subjectivité. Personnellement, je crois que les modèles d'analyse qui ont été présentés en dernière partie de la revue de littérature sont les plus susceptibles de conduire à des méthodes d'analyse d'items vraiment pertinentes. Je crois que les tests-critères constituent un type de mesure fondamentalement différent de celui des tests normatifs. Qu'on en juge seulement par les types de décision auxquels chacun doit nous conduire. Pour cette raison, j'ai tendance à me méfier des méthodes qui ne sont qu'une reformulation des méthodes utilisées pour les tests normatifs. Selon moi, le problème doit être abordé très différemment si on veut arriver à une solution originale et personnelle à ce type de test. Et si on débouche sur des méthodes qui ressemblent de très près à celles connues, ce pourra être tant mieux ou tant pis, mais le point de départ ayant été différent, nous pourrons avoir mieux confiance en la liberté de pensée que nous aurons prise envers les tests normatifs. Il ne faut pas chercher de fondement rationnel à cette opinion sauf que certaines théories de la créativité pourraient peut-être justifier cette approche.

Maintenant, regardons s'il est possible de comparer les divers types d'indices statistiques présentés. Comparons par exemple indice de difficulté et indice de sensibilité. Si on se réfère à Ivens, il devrait y avoir un changement maximal entre le niveau de difficulté au pré-test et le niveau de difficulté au post-test. Pour Roudabush, l'item le plus sensible à l'enseignement serait celui où la différence du rendement au pré-test et le rendement au post-test est maximal. Il y a donc de fortes similarités entre les deux indices et les deux devraient nous conduire à la rétention ou au rejet des mêmes indices. Une étude intéressante pourrait être effectuée pour comparer ces deux indices. L'indice de discrimination présente aussi une certaine similarité à ces deux indices, dans les cas où il est calculé pour discriminer le groupe qui n'a pas reçu d'enseignement

et celui qui a reçu un enseignement. Par contre, cet indice peut s'éloigner ou s'approcher beaucoup des deux premiers dépendamment de la méthode d'analyse qui aura été utilisée. Nous pouvons conclure que ces concepts qui de prime abord peuvent sembler bien différents peuvent conduire à des conclusions passablement similaires quant à la qualité de l'item.

Un point important doit ici être soulevé. Vu que souvent ces indices impliquent une comparaison pré-test et post-test, il est nécessaire de prendre en considération la qualité de l'enseignement effectué. Prenons le cas de l'indice de sensibilité. Si l'enseignement a été très efficace, il y aura beaucoup plus de possibilité de retrouver des valeurs élevées que si l'enseignement a été peu efficace. La valeur de l'indice dépend beaucoup de la qualité de l'enseignement et conséquemment il est difficile de savoir s'il faut attribuer la faiblesse d'un indice à l'item ou à l'enseignement.

Selon certains auteurs, dont Ivens, Hambleton et Gorth, il devrait y avoir homogénéité entre les items mesurant un même objectif. Pour eux, cette homogénéité se réflèterait dans des indices de difficulté équivalents pour chacun de ces items. Selon Dahl, les items mesurant un même objectif devraient porter des caractéristiques permettant de les distinguer de tous les autres groupements d'items. Ce par rapport à quoi ils devraient se distinguer n'est cependant pas clairement défini. Selon Kriewall, tous les items construits pour mesurer un même objectif devraient présenter la même probabilité de succès pour un individu donné. Nous constatons ainsi qu'avec des approches bien différentes, ces auteurs amènent tous l'idée d'homogénéité. Ils se distinguent cependant quant aux paramètres sur lesquels devraient se présenter cette homogénéité.

Je crois que ce concept d'homogénéité peut devenir un élément majeur dans les méthodes d'analyse d'items et que le travail de Dahl

sur le concept de congruence peut conduire à des éclaircissements importants dans ce domaine. Cependant, il faudra auparavant trouver réponse aux critiques portant contre ce concept. Selon Nitko, l'inverse, c'est-à-dire la non-homogénéité des items, pourrait même être défendue dans les cas d'objectifs pédagogiques sous-tendant la généralisation et le transfert de l'apprentissage. Selon moi, il s'agit d'un concept-clé sur lequel les auteurs devraient se pencher attentivement, ce qui apparemment n'a pas encore été fait.

En terminant, il nous parait nécessaire de soulever deux autres problèmes directement reliés au niveau de l'analyse d'items et dont nous n'avons pas fait mention ici. Il s'agit du nombre d'items à construire pour chacun des objectifs d'apprentissage mesurés et de la modalité ou format de ces items. Ces problèmes soulèvent d'importantes questions qu'il sera utile de solutionner concurremment aux méthodes d'analyse d'items. Tout ceci montre bien l'immense travail à accomplir par les chercheurs en docimologie pour arriver un jour à présenter aux constructeurs de test un volume auquel ceux-ci pourront se référer avec sécurité pour produire des tests qui répondent à toutes les exigences scientifiques.

BIBLIOGRAPHIE

AIRASIAN, Peter W., MADAUS, Georges F., Criterion-referenced testing in the classroom, in NCME Measurement in Education, Vol. 3, No. 4, May 1972, 8 p.

BESEL, Ronald, Using group performance to interpret individual responses to criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, New-Orleans, Lousiana, 1973, 10 p.

BESEL, Ronald, A Mastery-Learning Test Model, Presented at the 1972 annual American Educational Research Association, Chicago, Illinois, April 1972, 18 p.

BLOOM, B.S., HASTINGS, J.T., MADAUS, G.F., Summative evaluation, in Handbook on Formative and Summative Evaluation of Student Learning, New York, McGraw-Hill Book Company, 1971, pp. 61-138.

BORMUTH, J.R., On the Theory of Achievement Test Items, Chicago, University of Chicago Press, 1970, 163 p.

COX, Richard C., VARGAS, Julie S., A comparaison of item selection techniques for norm-referenced and criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, February 1966, 16 p.

DAHL, T., Toward an evaluative methodology for criterion-referenced measures: objective-item congruence. Paper delivered at the annual American Educational Research Association convention, San Drego, California, April 29 and 30, 1971.

DAVIS, F.B., Criterion referenced measurement, 1971, AERA Conference Summaries, ERIC, TM Report 12, 1972.

DAVIS, F.B., Criterion referenced measurement, 1972 AERA Conference Summaries, ERIC, TM Report 17, 1973, 6 p.

HAMBLETON, Leon P., RANDALL, Robert S., A model for estimating the reliability and validity of criterion- referenced measures, Paper presented at the annual meeting of the AERA, Illinois, April 1972, 21 p.

HAMBLETON, R.K., GORTH, W.P., Criterion-referenced testing: issues and applications. Center for Educational Research Technical Report # 13, School of Education, University of Massachusetts, Amherst, 1971, 27 pages.

HAMBLETON, Ronald K., NOVICK, Melvin R., Toward an integration of theory and method for criterion-referenced tests, in Journal of Educational Measurement, Vol. 10, No. 3, 1973, pp. 159-170.

HIVELY, W., DONLON, T., HOFMEISTER, A., ET AL., A bibliography on domain-referenced testing, For a workshop on domain-referenced testing in special education, University of Minnesota, Summer 1973, 5 p.

HIVELY, W., PATTERSON, H.L., PAGE, S., A "universe-defined" system of arithmetic achievement tests, in Journal of Educational Measurement, 1968, 5, pp. 275-290.

HSU, Tse-Chi, Empirical data on criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, New York City, 1971.

HSU, Tse-Chi, BOSTON, Elizabeth, Criterion-referenced measurement: an annotated bibliography. Learning Research and Development Center, University of Pittsburg, Feb. 1972, 20 p.

IVENS, S.H., An investigation of item analysis, reliability and validity in relation to criterion-referenced tests. Unpublished doctoral dissertation, Florida State University, 1970, 92 p.

IVENS, Stephen H., A pragmatic approach to criterion-referenced measures. Presented at annual American Educational Research Association and National Council on Measurement in Education, Chicago, Illinois, April 1972, 9 p.

JACKSON, Rex, Developing criterion-referenced tests. Princeton, New Jersey: ERIC, Clearinghouse on tests, Measurement and Evaluation, 1970, 18 p.

KELLER, Claudia Merhel, Criterion-referenced measurement: a bibliography. ERIC, Report No TM-R-No 7, 1972, 14 p.

KLEIN, Stephen P., KOSECOFF, Jacqueline, Issues and procedures in the development of criterion referenced tests. ERIC, TM Report 26, September 1973, 17 p.

KRIEWALL, T.E., Aspects and applications of criterion-referenced tests. Presented at the 1972 annual meeting of the American Educational Research Association, April 1972, 27 p.

KRIEWALL, Thomas E., HIRSCH, Edward, The development and interpretation of criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, California, Feb. 1969, 26 p.

LIVINGSTON, S.A., Criterion-referenced applications of classical test theory, in Journal of Educational Measurement, 1972, 9, pp. 13-26(a).

NITKO, A.J., A model for criterion-referenced test based on use. Paper presented at the annual meeting of the American Educational Research Association, New York, Feb. 1971, 16 p.

OAKLAND, Thomas, An evaluation of available models for estimating the reliability and validity of criterion referenced measures. Paper presented at American Educational Research Association, Illinois, April 1972, 7 p.

OSBURN, H.G., Item sampling for achievement testing, in Educational and Psychological Measurement, 28 (1968), pp. 95-104.

OZENNE, Dan G., Toward an evaluative methodology for criterion-referenced measures. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1972, 16 p.

POPHAM, W. James, Criterion-Referenced Measurement: An Introduction, Englewood Cliffs, New Jersey, Educational Technology Publications, 1971, 108 p.

POPHAM, W.J., HUSEK, T.R., Implications of criterion-referenced measurement, in Journal of Educational Measurement, 1969, 6, pp. 1-9.

RANDALL, Robert S., Contrasting norm referenced and criterion referenced measures. Prepared for American Educational Research Association symposium, Chicago, Illinois, April 1972, 10 p.

ROUDABUSH, G.E., Item selection for criterion referenced tests. Paper presented at annual American Educational Research Association meeting, New Orleans, 1973, 15 p.

ROUDABUSH, G.E., GREEN, D., Some reliability problems in a criterion-referenced test. Paper presented at the annual meeting of the American Educational Research Association, New York, 1971.

SMITH, Charles W., Criterion-referenced assessment. Paper presented at International Symposium on Educational Testing, The Netherlands, July 17, 1973, 14 p.

WOODCOCK, M.I., CHARLES, E., The issue of item and test variance for criterion-referenced tests. ERIC, TM 003 006, ED 079402, 5 p.